Abstract

Business management problems are often characterized by the availability of a large set of decision choices with a need to pick one or more from these choices in order to maximize the payoffs. Such problems present the dilemma on whether to explore the choices for improving the knowledge about the available choices or to exploit the currently available knowledge (on the choices) and pick the best choice based on the same. Since these Explore-Exploit tasks are opposing in nature, the decision maker has to obtain a trade-off between these two tasks. Multi-Armed Bandits (MAB), a family of Machine Learning algorithms, are tailor-made to handle such Explore-Exploit problem scenarios. The MAB problem is a sequential decision-making task where the decision maker (agent) decides to choose (pull), at each time step, an action (arm) from a pool of actions - based on some informed choosing strategy (policy). With the aim of maximizing the average payoff from this exercise in the long-run, the agent examines these payoffs to continuously improve the policy and decide on the future selection of arms. Alternatively, the same can be seen as a regret minimization problem where the regret is the difference between the rewards of an oracle policy that chooses the best arm in every time step and the rewards of the Agent's policy. In literature, it has been shown that in H pulls, $O(log\, H)$ regret is the lowest possible regret an MAB algorithm can achieve. MABs are studied in various settings and in this research we are interested in Stochastic MAB (SMAB) and Contextual MAB (CMAB). In a Stochastic MAB setting, each arm $i \in \{1, 2, \ldots, K\}$ is associated with

an unknown probability distribution v_i on $[0, 1]$ and rewards of this arm i which are independent and identically distributed ($i.i.d$) and are assumed to be drawn from that distribution v_i.

In this thesis, we propose effSAMWMIX, which achieves a logarithmic regret. effSAMWMIX's performance is compared with Thompson Sampling and KL-UCB algorithms over rewards which follow distributions like Exponential, Poisson, Normal distributions that are (suitably truncated over $[0, 1]$) along with Bernoulli, Triangular distributions. In addition, we performed experiments on these algorithms over a Synthetic distribution designed to stress test SMAB algorithms. We propose a variant of effSAMWMIX namely NBP-effSAMWMIX to address Online Portfolio Selection Problem (OPSP). OPSP has been tackled previously using a few machine learning approaches, including one that utilizes an SMAB as its decision engine and is referred to in the literature as Naive Bandit Portfolio (NBP) algorithm. An NBP's performance is expected to vary with the SMAB engine in the algorithm. As of now, only NBP-UCB1, which uses an established MAB algorithm named UCB1 as its kernel, is reported. We compare the performance of NBP-effSAMWMIX vis-à-vis NBP-KLUCB, NBP-TS and NBP-UCB1 algorithms. We tested the algorithms on both simulated and real-world market datasets and report the results. Further, we extend the effSAMWMIX to Ctx-effSAMWMIX, a CMAB which considers the additional contextual information. Availability of contextual information is common in business scenarios like a News Article Recommendations on a news website. These personalized recommendations play a major role in business success which could be a sale or a click on a news article in case of a news website. In News Article Recommendation problem, the aim is to choose a news article from an article list so that the shown article is of interest to target user. We model this problem as a CMAB. We employed an unbiased offline evaluation technique proposed in the literature to empirically test Ctx-effSAMWMIX

on Yahoo! Frontpage Today Module User Click Log R6B data set. The performance is measured on Click Through Rate (CTR), a measure that reports the number of clicks each recommended article obtained. Ctx-effSAMWMIX is compared with LinUCB which is cited well in CMAB literature.